SSURJET

Sir Syed University Research Journal of Engineering & Technology
2024, Vol. 14, No. 2,
https://doi.org/10.33317/ssurj.649

Optimizing Cardiovascular Risk Assessment with a Soft Voting Classifier Ensemble

Ammar Oad[1], Zulfikar Ahmed Maher[2], Imtiaz Hussain Koondhar[2], Karishima Kumari[3], and Hammad Bacha[4]

[1]Faculty of Information Engineering, Shaoyang University, Shaoyang, China
[2]Information Technology Center, Sindh Agriculture University, Tando Jam, Pakistan
[3]Isra University Hospital, Hyderabad, Pakistan
[4]Department of Emergency Medicine, The Second Xiangya Hospital, Central South University, Changsha, China

Correspondence Author: Ammar Oad (ammar.oad@hnsyu.edu.cn)

Received September 18, 2024; Revised December 02, 2024; Accepted December 09, 2024

Abstract

According to the latest data from the World Health Organization (WHO), heart disease has been the leading cause of death worldwide for the past several decades. It also includes a variety of heart-related disorders. Heart disease kills at least thirty individuals in Pakistan per hour. The best-known application of artificial intelligence is Machine Learning (ML). It is linked to numerous heart disease risk factors and the necessity of time to acquire sensitive accurate and dependable methods in order to make an early diagnosis. Experimental options have included the UCI repository's datasets on heart disease (which have 14 attributes) and cardiovascular diseases (12 attributes). The proposed ensemble soft voting classifier employs an ensemble of seven machine learning algorithms to provide binary classification; the Naïve Bayes, K-Nearest Neighbor, SVM Kernel, Decision Tree, Random Forest, Logistic Regression, and Support Vector Classifier. The accuracy, precision, recall, and F1_score value are provided by the suggested ensemble method with 70.9% 72.3% 68.6%, 70.1%, and Random Forest gives 71.5%, 72.2%, 70.3%, and 71.2%. The rest of the classifiers gave average scores. It means the proposed method provided the best results when compared with Decision Tree, Logistic Regression, Support Vector Classifier (SVC), SVM Kernel, K-Nearest Neighbor, and Naïve Bayes. Only Random Forest gives more accuracy than the proposed method on the cardio heart disease dataset.

Index Terms: *Cardiovascular Disease, Classifiers, Ensemble, Machine Learning, and Soft Voting.*

I. INTRODUCTION

Cardiovascular Disease (CVD) is Among the most common kinds of cause for death throughout the world. Women are generally less likely to develop CVD than men [1], but CVD diagnosis can be difficult because of the presence of various risk factors, such as high blood pressure, elevated cholesterol, and irregular heartbeat [2]. The cardiovascular system which consists of the heart and blood vessels may face many problems, including conduction defects, rheumatic heart disease, and endocarditis. These diseases are collectively called cardiovascular diseases [3]. Around the world, and particularly in high-income nations, coronary heart disease is thought to be the leading cause of mortality for women [4]. The World Health Organization (WHO) reported that in 2016, heart disease was responsible for 17.7 million deaths around the globe, accounting for approximately 31% deaths comprises of it. Moreover, according to the 2015 report by the World Health Organization (WHO), out of the 17 million premature deaths (below the age of 70) caused by non-communicable illnesses, 82 percent were in low- and middle-income nations, with cardiovascular illnesses or CVD accounting for 37% of the total (WHO 2016) [5]. These kinds of machine learning strategies are often applied to automatically learn hidden patterns in large data sets without human involvement [4]. To predict the membership function for labeling CVD data instances machine learning algorithms based on classification are employed. The process of classification involves extracting labels that describe significant classes of data [6]. Heart disease prediction performance is improved through utilizing group learning strategies [5]. In order to improve the outcomes, we consequently suggested using an ensemble approach. This paper is structured as follows: Section II goes over the literature in this field going into great detail about different machine learning and ensemble approaches. In Section III the suggested methodology is explained. In Section VI the suggested methodology analysis and results are covered. The suggested methodology outcomes are contrasted and examined with those of traditional machine learning algorithms.

II. RELATED WORK

In the last few years, there has been a remarkable amount of research enthusiastically embossed in the context of diabetes patient recognition using machine learning models and data mining operations. In 2021, Hana H. Alalawi, et al., [7] utilized deep gaining knowledge of networks and diverse devices gaining knowledge of class fashions to hit upon cardiovascular ailment the usage of two records units: Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Decision Tree (DT) Classifiers, Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF), Voting Classifier (VC), Gradient Boosting (GB) Classifier, and Naive Bayes (NB). On the all-inclusive cardiovascular dataset, the Gradient Boosting Classifier outperformed the Random Forest, which became the best classifier on the cardiovascular dataset. Authors utilized a device gaining knowledge of the approach to assemble an efficient computerized ailment diagnostic model for 3 critical diseases: Corona 19, heart ailment, and diabetes [7]. Following records access into the Android application, the proposed model employs a pre-skilled device, gaining knowledge of model skilled on the equal dataset and deployed on 'Firebase' to perform real-time database analysis and offer ailment analysis findings. Logistic Regression is utilized inside the Android application to do prediction computations. Coronavirus, heart ailment, and diabetes can all be averted with early identification. Here [8] suggested a cardiovascular ailment prediction model utilizing a device gaining knowledge of (ML) algorithm primarily based on the national medical health insurance agency's dataset. The strategies with the best average prediction accuracy were severe, Gradient Boosting, and Random Forest. The [9] used the Cleveland dataset with six extraordinary device-gaining knowledge of algorithms to expect the lifestyles of coronary artery ailment in patients. Researchers carried out 3 ML class modeling techniques to create a cardiovascular ailment detection model [10]. This attempt extracts the affected person's history leading to lethal heart ailment from a record set containing the affected person's clinical history, which includes chest discomfort, blood sugar, and blood stress, to forecast who has a cardiovascular ailment. The model is built on the usage of algorithms like Logistic Regression, Random Forest Classifier, and KNN. The effects showed that the KNN algorithm had the best accuracy of the 3. Some researchers employed device-gaining knowledge to figure out whether or no longer someone has a heart ailment [11]. K-Nearest Neighbor (KNN) and Random Forest are two supervised device-gaining knowledge of algorithms utilized in this work. K-Nearest Neighbor (KNN) has a prediction accuracy of 86.88%, whereas the Random Forest algorithm has an accuracy of 81.96% a class method was utilized to create the model, which plays an extensive part in prediction [12]. Logistic Regression, Random Forest, Support Vector Machine, Gaussian Naive Bayesian method, Gradient Boosting, K-Nearest Neighbor, Polynomial Naive Bayesian analysis, and Decision Tree. As a result, the Random Forest model is an effective and sensible method for figuring out heart contamination that can be utilized inside the clinical place an ensemble approach for a model class is constructed by combining Random Forest (RF), Naive Bayes (NB), and Gradient Boosting (GB) Classifiers [13]. They evaluated the performance of the suggested model by the usage of a real records set (70,000) accumulated from 'Kaggle' datasets. Inside the suggested model, the train and test records break up ratios are varied (50:50, 60: 40, 70:30, 80:20, and 87.5:12.5). The proposed model's AUC fee becomes 0.4 [14]; an ensemble approach for forecast cardiovascular disease and characteristic extraction algorithms like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), which utilize advanced techniques for analyzing coronary heart diseases. In the test, the packet ensemble schooling method using the DT and PCA characteristic extraction methods yielded first-rate results. In [15] the proposed RF trees, Naïve Bayesian SVMs, Neural Networks, and Logistic Regression-based Classifiers as an ensemble-based prototype version for predicting coronary heart ailment. They utilized this method to take a look at distinctive fact sets in the future because they constructed a generalized method. Handling diverse elegance labels in the prediction process can notably improve the effectiveness of fitness prognosis. Authors added in the prognosis of coronary heart ailment, a fee-powerful ensemble method based on 5 distinctive classifiers [16]. To decide how powerful, the ensemble classifier becomes, it will be tested at the coronary heart ailment Statlog dataset, the Cleveland heart disease dataset, and the Hungarian heart ailment dataset. Then checked out E, MC, Gmean, accuracy, recollect, specificity, and AUC. The reliability of the results is checked using Wilcoxon mark ranking criteria. The results confirmed that the proposed method can offer promising results for the prognosis of coronary heart disease in comparison with man or woman classifiers and previous research. A group of researchers suggested a more advantageous supervised learning method for forecasting the hazard of coronary heart disease, using an average-based partitioning strategy, this method randomly walls facts set into smaller businesses [17]. Experimental results were completed with classification accuracy of 93% and 91% on the Cleveland and Framingham datasets, respectively. The advanced overall performance of the suggested ensemble training method is in addition shown by using the receiver's overall performance curve. The findings propose that the proposed ensemble can also as it should be predicting coronary. The [18] utilized an ensemble method to take a look at the prediction for cardiovascular ailment based on the facts obtained. The Statlog coronary heart disease become utilized because of the facts set. The effectiveness of the suggested strategy is assessed using overall performance criteria along with accuracy, sensitivity, and specificity. The proposed method has proven to be 87.04% correct. In an assessment of different research published on the UCI website, they deemed it to be one of the first-rate based on the results gathered. In [19] authors constructed an ensemble supervised learning method, utilizing a supervised learning classifier, to enhance the accuracy of a supervised learning algorithm using Python with Scikit getting to know. For predicting coronary heart disease, algorithms along with Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, and K-Nearest Neighbor have been tested, with Logistic Regression outperforming the results with an accuracy of 86.81%. The methods of ensemble bagging, boosting, and voting were investigated. The vote classifier method had the very best accuracy at the check set, at 96.55%. Reference [20] suggested an ensemble version based on a Random Forest set of rules and Support Vectors with the UCI system Learning Repository. The results claimed to have an accuracy of 89% in detecting coronary heart disease. The [21] suggested Maintenance Decision Support System (MDSS) as a tool for predicting the prognosis of atherosclerosis in its early tiers. The suggested gadget generates capabilities adequate for predicting sufferers with or without atherosclerotic ailment using 3 systems-Supervised learning algorithms (ANN, AdaBoost, and DT algorithms) for the facts set. Using the clinical facts set, 835 samples were collected from databases in Cleveland, Hungary, and Zalizade Sani. The suggested method has the very best accuracy of 94%. In [22] researchers created a gadget that could recognize regulations that forecast a patient's hazard degree based on a fixed of fitness indicators, particularly coronary heart disease to identify hidden styles and predict whether patients have coronary heart disease. To test diverse facts mining strategies, researchers are using the Cleveland coronary heart disease dataset from the University of California, Irvine (UCI) Machine Learning Repository. As a result, the Logistical Regression set of rules becomes determined to be the simplest, most of the 4 algorithms with an accuracy of 82.89%. The reference [23], applied diverse capabilities of cardiac ailment on supervised learning algorithms such as Naive Bayes, Decision Tree, Nearest Neighbor, and Random Forest set of rules. The results show that using the Nearest Neighbor yielded the very best accuracy rating. [24] and [25].

III. PROPOSED METHODOLOGY

The proposed ensemble tender balloting classifier uses the ensemble of seven system mastering algorithms viz. choice tree, Random forest, Logistic regression, assist Vector Classifier (SVC), SVM Kernel, K-Nearest Neighbor and Naïve Bayes and will be trained with Cleveland database of UCI repository of coronary heart disease dataset with 14 attributes [5] and another dataset with twelve attributes of Cardiovascular illnesses [6] for classifying the presence of coronary heart sickness in patients. The first dataset has 76 characteristics and 303 occurrences. The cardiovascular disease dataset has 70,000 patient information entries with 11 features, although only 14 of the 76 attributes were examined. There are 35,021 facts for Cardio 0 patients and 34,979 facts for Cardio 1 patients, making up 70,000 different facts. The aim of this study is to extend the ensemble method and to compare it with other traditional methods to improve the prediction of cardiac patients.

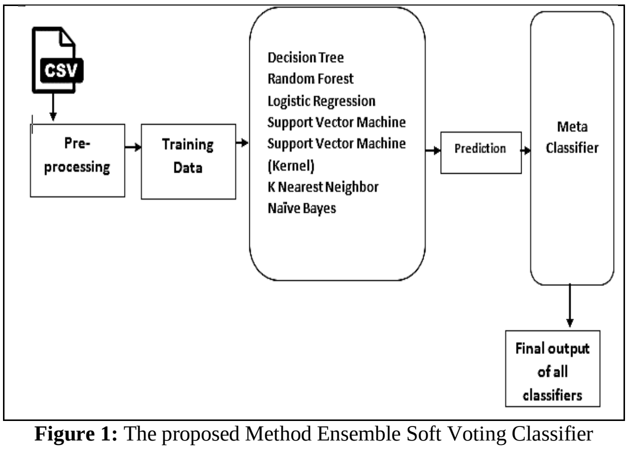

Figure 1: The proposed Method Ensemble Soft Voting Classifier

Figure 1 demonstrates the proposed technique, where a dataset is given and applied pre-processing and trained with the proposed technique and gets results of every classifier, and finally apply Soft Voting Classifier to predict the final output.

A. Algorithm

1: Procedure choose (Dataset1)
2: Select Features and Target of Dataset1
3: Procedure partitioned_data (Dataset1)
train_set, test_set=split (attributes, label)
Return train_set, test_set
Voting="soft"
CL 1 = Logistic Regression (train_set, test_labels, test_set,)
CL 2 = RandomForestClassifier (train_set, test_labels, test_set)
CL 3 = tree.DecisionTreeClassifier (train_set, test_labels, test_set)
CL 4 = SVM (train_set, test_labels, test_set)
CL 5 = SVM_Kernel (Training_data, Testing_labels, Testing_data)
CL 6 = K_Neighbors (Training_data, Testing_labels, Testing_data)
CL 7 =Naïve_Bayes (Training_data, Testing_labels, Testing_data)
S_V_CLASSIFIER=concatenate (CL1, CL2, CL3, CL4, CL5, CL6, CL7)
S_V_CLASSIFIER.FIT (Training_data, Testing_labels, Testing_data)
Predictions = S_V_CLASSIFIER.predict (Testing_data)

B. Data Description

In this study's paintings, the Cleveland database of UCI repository of heart disorder dataset includes 12 attributes, and any other dataset with 14 attributes of Cardiovascular diseases was taken into consideration for classifying the presence of heart disorder in sufferers. The primary dataset contains 303 instances and 76 attributes, from these 76 attributes, for testing the 12 best-performing features are considered, and another cardiovascular dataset contains 70,000 patient records using 11 features out of 70,000 records, 35,021 patient records with 0 aerobic, and 34,979 patient records with 0 aerobic.

1) Dataset 1:

The patient information in the Heart UCI dataset contains a target variable that indicates whether heart disease is present or absent. This dataset consists of 76 attributes, but the most effective 12 are applied in our assessments to ensure that our effects are comparable to those in previous gadget studying papers. The chosen developments and their homes are proven in desk 1. The entire wide variety of information on this dataset is 303, which can be a bit wide variety whilst compared to the hundreds to hundreds of hundreds of facts focuses in an everyday gadget studying dataset. A few considers have seemed into the execution of various gadget studying calculations in this dataset. Considering the fact that of the dataset's ubiquity, it is straightforward to peer how competitive the graduate students come about are, in addition to how the AutoML approach compares to human–specialists' frameworks. On this dataset, the goal variable is ' goal' in desk 1. 138 information is for sufferers with a goal and 165 information is for sufferers with goal 1 (see Table I).

Table I: Detailed Overview of the UCI Heart Dataset's Twelve Variables

S. No.	Attribute Name	Description
1	Age	Age Measured in Days
2	Height	Height in centimeters
3	Weight	Weight in kilograms
4	Gender	Gender (1 = Female; 2 = Male)
5	Ap_hi	Systolic Blood Pressure (in mm Hg)
6	Ap_lo	Diastolic Blood Pressure (in mm Hg)
7	Cholesterol	Cholesterol Levels in Serum (1 = Normal; 2 = Elevated; 3 = Significantly Elevated)

8	Gluc	Glucose Levels (1 = Normal; 2 = Elevated; 3 = Significantly Elevated)
9	Smoke	Smoking Status (0 = No; 1 = Yes)
10	Alco	Alcohol Consumption (0 = No; 1 = Yes)
11	Active	Exercise Habits (0 = No; 1 = Yes)
12	Cardio	Presence of Cardiovascular Disease (0 = No; 1 = Yes)

2) Dataset 2:

The cardiovascular dataset contains 70,000 silent data, and the target (aerobic) characterizes the presence or absence of cardiac contamination using 14 indicators (see Table II). There are three kinds of input highlights: objective (contains records), exam (contains the outcomes of a healing exam), and subjective (contains the outcomes of a restorative exam) (contains records given with the aid of the persistent). In this dataset, the target variable is ' aerobic' in desk 2. 35,021 information of sufferers with aerobic and 34,979 information of sufferers with aerobic 1 are a number of the 70,000 information.

Table II: Fourteen of Cardiovascular Disease Dataset's Fourteen Variables

S. No.	Type of Attribute	Description
1	Age	Age in Years
2	Gender	Gender (1 = Male; 0 = Female)
3	CP	Discrete: Type of Chest Pain, Classified into 4 Categories
4	Trestbps	Blood Pressure During Hospital Admission in mm Hg
5	Chol	Cholesterol Level in mg/dl
6	FBS	Fasting Blood Sugar Level (1 = Above 120 mg/dl, 0 = Below)
7	Restecg	Resting Electrocardiogram Results (Values: 0, 1, 2)
8	Talah	Maximum Heart Rate during Exercise
9	Exang	Presence of Angina (1 = Yes, 0 = No)
10	Oldpeak	Depression of ST-Segment during Exercise Compared to Rest
11	Slope	Discrete: Slope of the ST Segment during Exercise (Values: 0, 1, 2)
12	CA	Discrete: Number of Major Coronary Vessels (0-4) observed during Fluoroscopy
13	Thalach	Type of Defect (Values: 0-3)
14	Cardio	Presence of Cardiovascular Disease (1 = Present, 0 = Absent)

C. Data Pre-Processing

Information Pre-processing may be a pivotal step that changes information right into a usable and effective arrange that may be strengthened to a machine-learning calculation. Information normalization is the primary strategy utilized for information pre-processing. This strategy is utilized to carry out direct information changes. It is miles furthermore known as Min-max normalization because the attributes' entire values drop between [0, 1]. Name encoding is the following pre-processing strategy utilized. This strategy is utilized at the subordinate variable, that's whether or not the man or woman has diabetes. As a result, all of the string values within the yield variable are supplanted with and 1, coming about within the yield course being decided.

IV. RESULTS AND DISCUSSION

An ensemble method with seven classifiers has been developed with Python language, actually, this method was designed and developed to predict heart disease in patients. The classifiers that have been added to an ensemble method are Logistic Regression, Random Forest, Decision Tree, SVC, SVC (Kernel), K-Neighbor, and Naïve Bayes. These classifiers and the proposed ensemble method have been evaluated with cross-validation accuracy and with Confusion Matrix, F1, precision, recall, and accuracy. Two datasets were selected, Dataset 1 with 12 attributes and 303 patient records shown in Table I, and Dataset 2 with 14 attributes and 70,000 records of patients shown in Table II. 70% data was selected for training purposes and 30% for testing purposes. As a result, it was found that:

A. Dataset 1

The Logistic Regression, Soft Voting (proposed method), Naïve Bayes, and Decision Tree Classifiers provided the highest F1, precision, recall, and accuracy scores and the rest of the classifiers provided average scores for Confusion Matrix. For cross-validation, the Random forest and Soft Voting Classifier yielded 84% accuracy, while Logistic Regression and Naïve Bayes yielded 85% and the rest of the classifier provided average accuracy shown in Table III, Figure 2(A), and Figure 3(A). The overall performance of all classification models for Dataset 1 has been represented using the ROC Curve that shows 96% for Random Forest, 93% for Logistic Regression, and 92% for Naïve Bayes and Soft Voting Classifier as shown in Figure 4.

Table III: Cross Validation Accuracy, F1, Recall, Precision and Accuracy Scores of classifiers with Dataset1

S. No.	Classifier	Cross Validation Accuracy	F1 Score	Recall Score	Precision Score
1	Logistic Regression	70	69	67	71
2	Random Forest	72	71	70	72
3	Decision Tree	63	64	64	63
4	Support Vector Machine	60	58	56	61
5	SVM (Kernel type =linear)	60	58	56	61
6	K-Nearest Neighbor (KNN)	63	62	62	62
7	Naive Bayes	59	44	32	71
8	Soft Voting Classifier	70	70	68	72

B. Dataset 2

The Random Forest provided the highest score for F1 which is 71%, after that the Soft Voting Classifier provided 70 % and then Logistic Regression provided 69% rest of the classifiers provided average F1 scores for Confusion Matrix. In the cross-validation, the Random Forest provided 72% accuracy, and Soft Voting and Logistic Regression provided 70% accuracy, the rest of the classifiers provided average scores for cross-validation and for Confusion Matrix, F1, precision, recall, and accuracy, shown in Table IV, Figures: 2(B), 3(B), and 5. This Table shows cross-validation accuracy, F1, recall, precision, and accuracy The results of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, SVM Kernel, and Logistic Regression Soft Voting Classifier.

Table IV: Cross Validation Accuracy, F1, Recall, Precision and Accuracy Scores of classifiers with Dataset2

S. No.	Classifier	Cross Validation Accuracy	F1 Score	Recall Score	Precision Score
1	Logistic Regression	85	92	90	93
2	Random Forest	84	88	84	93
3	Decision Tree	79	79	81	77
4	Support Vector Machine	64	71	93	57
5	SVM (Kernel type =linear)	64	71	93	57
6	K-Nearest Neighbor (KNN)	60	51	48	55
7	Naive Bayes	85	87	87	87
8	Soft Voting Classifier	84	89	90	88

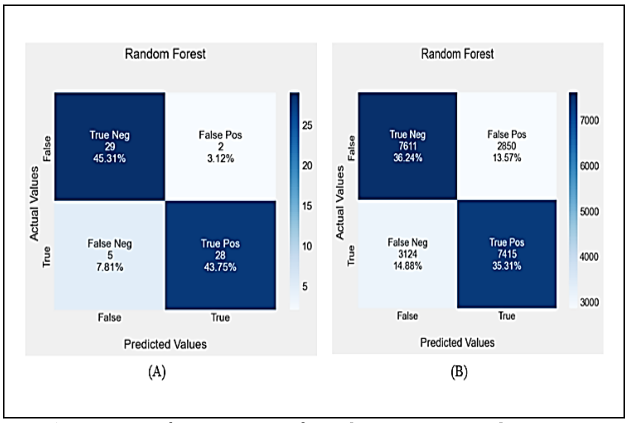

Figure 2: Confusion Matrix of Random Forest in Both Datasets

Figure 2(A) shows Cleveland Database of UCI Repository of Heart Disease Dataset: This figure shows Confusion Matrix score of True Positive score is 43.75%, 45.31% True Negative, 3.12% False Positive, and, 7.81% False Negative. Figure 2(B) shows Cardiovascular Disease Dataset: This figure shows the True Positive score that is 35.31%, 36.24% True Negative, 13.57% False Positive, and, 14.88% False Negative.

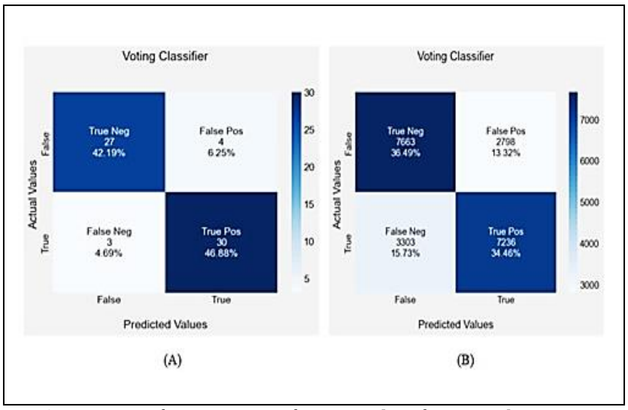

Figure 3: Confusion Matrix of Voting Classifier in Both Datasets

Figure 3(A) shows Cleveland Database of UCI Repository of Heart Disease Dataset: This figure represents the True Positive score is 46.88%, 42.19% True Negative, 6.25% False Positive, and, 4.69% False Negative.

Figure 3(B) shows Cardiovascular Disease Dataset: This figure shows the True Positive score is 34.46%, 36.49% True Negative, 13.32% False Positive, and 15.73% False Negative.

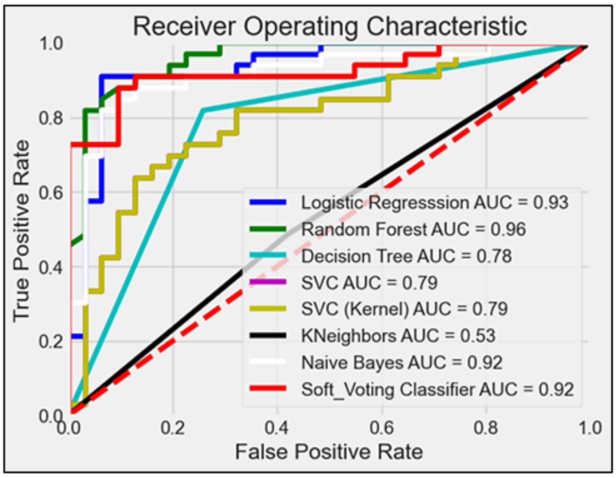

Figure 4: The ROC Curve of all Classifiers with Dataset1

This is a ROC Curve of Logistic Regression, Random Forest, Decision Tree, SVC, SVC Kernel, K-Nearest, Naïve Bayes, and Soft Voting Classifier. This shows that the highest area is given to Random Forest which is 0.96, after that 0.93 is Logistic Regression, after that Naïve Bayes and Soft Voting Classifiers have 0.92 area, and the rest of the classifiers have less percent (see Figure 4).

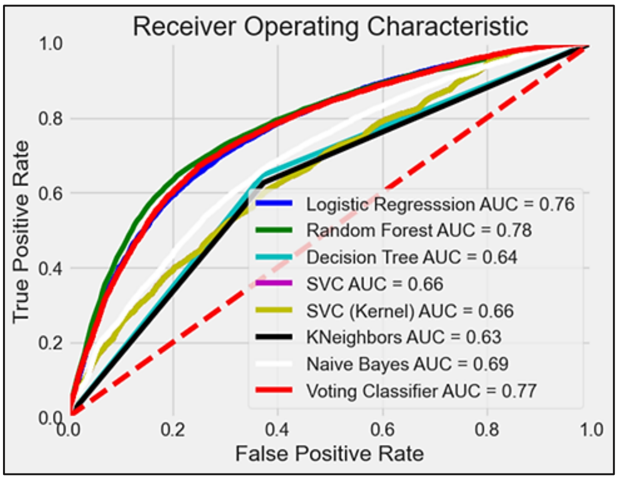

Figure 5: The ROC Curve of all Classifiers with Dataset2

This is an ROC Curve of Logistic Regression, Random Forest, Decision Tree, SVC, SVC Kernel, K-Nearest Neighbors, Naïve Bayes, and Soft Voting Classifier. This shows that the highest area is given to Random Forest which is 0.78, after that 0.77 is for Soft Voting Classifier and the rest of the classifiers have less percent (see Figure 5).

V. CONCLUSION

It is concluded Random Forest has the highest accuracy which is 96% in Dataset 1 and 78% in Dataset 2. The Soft Voting Classifier has got 92% accuracy in Dataset 1 and 77% accuracy in Dataset 2 the Logistic Regression has got 93% accuracy in Dataset 1 and 76% accuracy in Dataset 2. Moreover, all other classifiers have average accuracy. This means that Random Forest performed best in all classifiers and the proposed Soft Voting Classifier is at second number. So this has been proved that the Soft Voting Classifier can predict the best results excluding Random Forest. Our ensemble method shows competitive performance, but recent CNN approaches consistently surpass 90% accuracy. In the future, the Soft Voting Classifier may be added to AdaBoost (Adaptive Boosting) or another technique to improve its prediction. Such ensemble techniques can be applied for marketing research, health issues for medical research, agriculture research, pharmacology research, etc., for better accuracy in future decisions.

Acknowledgment

Our research group would like to thank everyone who has contributed to the successful completion of this work and Muneer Ahmed Rajper for his infrastructural support.

Authors Contributions

All the authors equally conducted and contributed to this study.

Conflict of Interest

There is no conflict of interest between all the authors.

Data Availability Statement

The testing data is available in this paper.

Funding

This research received no external funding.

References

[1] Gao, Z., Chen, Z., Sun, A., & Deng, X. (2019). Gender Differences in Cardiovascular Disease. *Medicine in Novel Technology and Devices, 4*, 100025.

[2] Nadakinamani, R. G., Reyana, A., Kautish, S., Vibith, A. S., Gupta, Y., Abdelwahab, S. F., & Mohamed, A. W. (2022). [Retracted] Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. *Computational intelligence and neuroscience, 2022*(1), 2973324.

[3] Lopez, E. O., Ballard, B. D., & Jan, A. (2023). Cardiovascular Disease. In *StatPearls [Internet]*. StatPearls Publishing.

[4] Woodward, M. (2019). Cardiovascular Disease and the Female Disadvantage. *International Journal Of Environmental Research and Public Health, 16*(7), 1165.

[5] UC Irvine. (n.d.). Dataset1. Heart Disease. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[6] Kaggle.com. (n.d.). Dataset2. Cardiovascular Disease Dataset. Retrieved from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv

[7] Alalawi, H. H., & Alsuwat, M. S. (2021). Detection of Cardiovascular Disease Using Machine Learning Classification Models. *International Journal of Engineering Research & Technology, 10*(7), 151-7.

[8] Kumar, N., Narayan Das, N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient Automated Disease Diagnosis Using Machine Learning Models. *Journal of Healthcare Engineering, 2021*(1), 9983652.

[9] Kim, J. O., Jeong, Y. S., Kim, J. H., Lee, J. W., Park, D., & Kim, H. S. (2021). Machine Learning-Based Cardiovascular Disease Prediction Model: A Cohort Study on the Korean National Health Insurance Service Health Screening Database. *Diagnostics, 11*(6), 943.

[10] Akella, A., & Akella, S. (2021). Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward An Open Source Solution. *Future Science OA, 7*(6), FSO698.

[11] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart Disease Prediction Using Machine Learning Algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[12] Garg, A., Sharma, B., & Khan, R. (2021). Heart Disease Prediction Using Machine Learning Techniques. In *IOP Conference series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.

[13] Padmaja, B., Srinidhi, C., Sindhu, K., Vanaja, K., Deepika, N. M., & Patro, E. K. R. (2021). Early and Accurate Prediction of Heart Disease Using Machine Learning Model (Institute of Aeronautical Engineering, Hyderabad, India). *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(6), 4516-4528.

[14] Uddin, M. N., & Halder, R. K. (2021). An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach. *Informatics in Medicine Unlocked, 24*, 100584.

[15] Gao, X. Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity, 2021*(1), 6663455.

[16] Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble Approach for Developing A Smart Heart Disease Prediction System Using Classification Algorithms. *Research Reports in Clinical Cardiology*, 33-45.

[17] Zhenya, Q., & Zhang, Z. (2021). A Hybrid Cost-Sensitive Ensemble For Heart Disease Prediction. *BMC Medical Informatics And Decision Making, 21*, 1-18.

[18] Mienye, I. D., Sun, Y., & Wang, Z. (2020). An Improved Ensemble Learning Approach For The Prediction of Heart Disease Risk. *Informatics in Medicine Unlocked, 20*, 100402.

[19] Emakhu, J., Shrestha, S., & Arslanturk, S. (2020, August). Prediction System for Heart Disease Based on Ensemble Classifiers. In *The 5th NA International Conference on Industrial Engineering and Operations Management Detroit* (pp. 10-14).

[20] RutujaGujare, D. V., & Bhatt, S. (2020). Enhanced Heart Disease Prediction Using Ensemble Learning Methods. *International Journal of Advanced Science and Technology, 29*(06).

[21] Kashif, F., & Yusof, U. K. (2020, December). Detection of Cardiovascular Disease Using Ensemble Machine Learning Techniques. In *International Conference of Reliable Information and Communication Technology* (pp. 176-185). Cham: Springer International Publishing.

[22] Terrada, O., Hamida, S., Cherradi, B., Raihani, A., & Bouattane, O. (2020). Supervised Machine Learning Based Medical Diagnosis Support System For Prediction of Patients with Heart Disease. *Advances in Science, Technology and Engineering Systems Journal, 5*(5), 269-277.

[23] Magar, R., Memane, R., Raut, R. (2020).Heart Disease Prediction Using Machine Learning. *Journal of Emerging Technologies and Innovative Research, 7*(6).

[25] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction Using Machine Learning Techniques. *SN Computer Science, 1*(6), 345.

ABBREVIATIONS AND ACRONYMS

- CVD Cardio Vascular Disease
- AHA American Heart Association
- AI Artificial Intelligence
- ML Machine Learning
- PAD Peripheral arterial disease
- WHO World Health Organization
- CVD Cardio Vascular Disease
- VC Voting Classifier
- LR Logistic Regression
- MLDS Multilayer Dynamic System
- CAE Correlation Attribute Estimator
- IGAE Information Acquisition Attribute Estimator
- GRAE Gain Attribute Estimator
- CART Classification and Regression Trees
- UCI University of California
- CHD Cleveland Heart Disease